Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT

Year: 2022 | Citations: 207 | Authors: Thilo Hagendorff, Sarah Fabi, Michal Kosinski

---

Abstract

We design a battery of semantic illusions and cognitive reflection tests, aimed to elicit intuitive yet erroneous responses. We administer these tasks, traditionally used to study reasoning and decision-making in humans, to OpenAI's generative pre-trained transformer model family. The results show that as the models expand in size and linguistic proficiency they increasingly display human-like intuitive system 1 thinking and associated cognitive errors. This pattern shifts notably with the introduction of ChatGPT models, which tend to respond correctly, avoiding the traps embedded in the tasks. Both ChatGPT-3.5 and 4 utilize the input–output context window to engage in chain-of-thought reasoning, reminiscent of how people use notepads to support their system 2 thinking. Yet, they remain accurate even when prevented from engaging in chain-of-thought reasoning, indicating that their system-1-like next-word generation processes are more accurate than those of older models. Our findings highlight the value of applying psychological methodologies to study large language models, as this can uncover previously undetected emergent characteristics. The reasoning capabilities of OpenAI's generative pre-trained transformer family were tested using semantic illusions and cognitive reflection tests that are typically used in human studies. While early models were prone to human-like cognitive errors, ChatGPT decisively outperformed humans, avoiding the cognitive traps embedded in the tasks.